ESTIMATION AND VARIANCE ESTIMATION IN A STANDARD ECONOMIC PROCESSING SYSTEM

Richard Sigman[1], U.S. Census Bureau
ESMPD, Room 3108-4, U.S. Census Bureau, Washington DC 20233, USA rsigman@census.gov

The U.S. Census Bureau has developed software called the Standardized Economic Processing System, or StEPS, that it plans to use to replace 16 separate systems, which are currently used to process over 100 current economic surveys. This paper describes the methodology and design of the StEPS modules for estimation and variance estimation and chronicles our experiences in using these modules to migrate surveys into StEPS. The paper concludes with a discussion of possible future enhancements to the estimation and variance estimation functions in StEPS.

Key Words: Survey Processing, Economic Surveys, StEPS

1. Introduction

The U.S. Census Bureau conducts over one hundred establishment surveys. Many of these are surveys of commercial businesses. A small number are surveys of government institutions. The Census Bureau refers to these surveys as *economic surveys* because they collect quantitative data describing business operations of survey units. Also, these surveys provide economists and other analysts with estimates and data sets needed for macro- and micro-economic analyses. For example, the Bureau of Economic Analysis uses estimates from economic surveys to determine the national income and expense accounts.

Economic surveys can differ widely with respect to characteristics of reporting units and content of survey questions. They are often similar, however, with respect to data-processing requirements, which has prompted the Census Bureau to begin consolidating the survey-processing systems for many of its economic surveys. The development and use of generalized software, called the Standard Economic Processing System (StEPS) has made this possible.

This paper describes the current capabilities of StEPS for calculating survey estimates and associated sampling errors. Sections two through four provide background material. In particular, section two summarizes the characteristics of Census Bureau economic surveys that are relevant to calculating survey estimates and sampling variances. Section three discusses variance estimation methods: those used in the legacy systems, those evaluated for StEPS, and those currently implemented in StEPS. Section four briefly describes the entire StEPS system. Sections five through eight focus on the StEPS Estimates and Variances Module. Section five describes the components of the module, and section six presents and explains two examples of StEPS estimation scripts. Section seven describes our implementation experiences for the Estimates and Variances Module in 1998 and 1999. Finally, section eight describes future activities and possible enhancements.

2. Economic Surveys Conducted by the Census Bureau

The Census Bureau consists of several directorates that conduct censuses and surveys. The most widely known directorate is the Decennial Census Directorate, which conducts the demographic decennial census. Another directorate, called the Economic Programs Directorate, conducts economic censuses every five years and conducts current economic surveys monthly, quarterly, and annually in areas of manufacturing, construction, commercial services, government services, and foreign trade. The directorates, such as the Decennial Census Directorate and the Economic Programs Directorate, are responsible for developing survey methods and associated processing systems for the censuses and surveys they conduct.

In 1995 the Economic Directorate began to consolidate its processing systems for current surveys. This was preceded, however, by two activities that provided information that was used in planning and directing the consolidation effort.

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform parties and to encourage discussion.

One of these was a compilation of an inventory of the Economic Directorate's statistical practices carried out by King and Kornbau (1994). Some of their findings (along with updates for recent survey changes) are the following:

- Sample designs are primarily single stage, but a number of different sampling methods are used: stratified sampling, cut-off sampling, Poisson sampling, and Tillé sampling.
- Only four types of estimators for (unraked) totals are used: unweighted estimator (used by cut-off surveys), Horvitz-Thompson estimator, ratio estimators (combined and separate), and link-relative estimator. (As a result of a 1997 redesign, the monthly surveys of retail and wholesale trade no longer use composite estimation.)
- A number of different variance estimation methods are used by the economic surveys that calculate sampling variances: jackknife, method of random groups, balanced repeated replication and sampling-theory formulas.

The other activity that provided useful information was the development in 1994 of a processing system for the Farm and Ranch Irrigation Survey (FRIS). This demonstrated the feasibility of the following in developing a production processing system for an economic survey:

- The use of reusable SAS® code configured to individual surveys by analyst-specified parameters,
- The use of interactive screens to allow analysts to specify parameters, and
- The use of a general-purpose variance-estimation program, VPLX, to allow sample designers to specify how to calculate standard errors.

3. Variance Estimation Methods

The Economic Directorate decided to consolidate its current-survey processing systems by developing a Standard Economic Processing System, called StEPS. Because one of the functional requirements for StEPS was the estimation of sampling variances, we carried out several research studies on variance estimation during the development of StEPS. These studies explored two possible development approaches: (1) reduce the number of different variance estimation methods, and (2) use available computer programs for calculating design-based variances.

The available computer programs we studied were VPLX, developed in-house by Fay (1990), and SUDAAN, developed by Research Triangle Institute (RTI). VPLX estimates variances by means of replication . It contains options to estimate variances using random groups, jackknife, stratified jackknife, balanced repeated replication, and generalized replication. (See Wolter 1985 or Rust 1985 for descriptions of these variance estimation methods.) Prior to the use of VPLX in the FRIS system, VPLX had been used very little by the Census Bureau's economic surveys, but it is used extensively by the Census Bureau's demographic surveys of households and by the 2000 population census. RTI has recently added replication-based methods to SUDAAN, but at the time of our investigation these methods were not available.

By performing repeated stratified sampling from a simulated FRIS population, Tremblay and Sigman (1996) compared stratified-jackknife variance estimates calculated by VPLX to variances estimates calculated by SUDAAN using sampling theory formulas--i.e., S^2 formulas with Taylor-series approximations. One objective of this study was to evaluate the two programs as to their suitability for inclusion in StEPS. A second objective, however, was to determine if in stratified-sample designs the use of the stratified jackknife could replace the use of sampling-theory formulas, which would be more difficult to program compared to the stratified jackknife. As expected, Tremblay and Sigman found that for linear estimates, the two programs/methods yielded identical results. They found, however, that "[f]or separate-ratio estimation, the larger absolute bias of SUDAAN and the larger variance of VPLX tend to balance each other when one considers the root mean square errors of calculated standard errors." Tremblay and Sigman concluded that the choice between SUDAAN or VPLX was not obvious in terms of studied statistical properties. They recommended that StEPS use VPLX, however, for the following reasons: "VPLX is more flexible: basically, anything that can be set up in a formula can be done in VPLX. ... VPLX is 'license free', and consulting is more readily available since its developer/maintainer is resident at the Census Bureau."

Tremblay (1996) extended Tremblay and Sigman (1996) by using VPLX to additionally calculate variances using the method of random groups with 16 groups. She compared the SUDAAN results (i.e.,S^2 formulas with Taylor-series approximations), the VPLX stratified-jackknife results, and the VPLX-random-group results for two different FRIS survey variables and found that in nearly all cases the estimated relative root-mean-square errors of the VPLX-random-group estimated standard errors were larger than both those from SUDAAN and those from the VPLX stratified jackknife. This difference was particularly pronounced for the case of aggregated multi-stratum estimates.

The Economic Directorate uses the method of random groups to estimate variances for its monthly and annual surveys of retail and wholesale trade and for its annual surveys of other service industries. Rust (1985) states the "[t]he random groups method is most useful in surveys using a large number of PSUs, where either many PSUs are selected per stratum, or few gains are believed to result from the finer levels of stratification." The surveys for which the Economic Directorate uses the random groups method to estimate sampling variances satisfy these requirements. Town (1997) used the random groups option of VPLX to calculate sampling variances for two surveys: FRIS and the Annual Trade Survey (ATS) for wholesale trade. She found that the number of lines of VPLX code that were required to calculate variances for these two surveys were very different. For FRIS only 77 lines of VPLX code were required, whereas for ATS 3591 lines of VPLX code were needed. Some of the reasons for this difference were the following:

! The VPLX program for ATS calculated variances for 17 different types of estimates: 7 type of unadjusted estimates and 10 types of census-adjusted estimates. These included raked and unraked estimates of level, ratios, percentages, trends of level estimates, and trends of ratios. The VPLX programs for FRIS, on the other hand, calculated variances for only one type of estimate: census-adjusted estimates of level.

! The VPLX program for ATS calculated variances for 22,309 estimates; whereas the VPLX program for FRIS calculated variances for only 276 estimates.

! The VPLX program for ATS calculated a number of derived items, whereas the VPLX program for FRIS did not.

At the time of Town's investigation some enhancements to the syntax rules for VPLX code had been completed and more were planned for the future. Dr. Robert Fay, the developer of VPLX, recoded a portion of Town's VPLX program for ATS using the proposed enhancements to the VPLX syntax rules. Using the proposed syntax rules, the VPLX code was 152 lines; using the proposed new syntax rules, however, it was only 31 lines. From her two comparisons--FRIS versus ATS and "old" syntax versus "new" syntax--Town concluded that VPLX could be used by StEPS to calculate variances for surveys that use the random groups method, but the implementation effort of using VPLX for surveys like ATS appeared to be quite high. Town recommended that the new syntax rules for VPLX be used when they become available and that StEPS developers also "investigate alternative software approaches, such as the calculation by StEPS of random-group-level estimates followed by variance estimation using the method of random groups performed by a general-purpose SAS macro."

Dajani (1999) further explored the random group method of variance estimation in order to make recommendations on how it should be used in StEPS. Dajani studied the problem of how the method of random groups should be used to estimate variances for aggregate estimates following the estimation of variances for more detailed estimates. This same problem was studied by Kott (1999) in the context of using the delete-a-group jackknife to estimate sampling variances. In Kott's study the aggregate estimates were national-level estimates, and the detailed estimates were state-level estimates; whereas in Dajani's study the aggregate estimates were for two- and three-digit Standard Industrial Codes (SICs) and the detailed estimates were for four-digit SICs. One approach to estimating the variances for the aggregate estimates is to use the same replication method for the aggregate estimates as was used for the detailed estimates. A second approach, labeled the *hybrid method* by Kott, is for linear estimators to sum the variances of the detailed estimates to the aggregate level. For non-linear estimators one estimates the variance of a first-order Taylor-series approximation to the aggregate estimator, which is a linear combination of variances and covariances of the aggregate totals, calculated by summation of the variances and covariances of detail totals.

Like Town, Dajani studied ATS; but Town used 1995 ATS data (the sample for which was selected in 1990), whereas Dajani used 1997 ATS data (the sample for which selected in 1995). Dajani compared the two different approaches for estimating the variances of aggregate estimates for ATS aggregate totals, ratios of aggregate totals, and trends of aggregate totals. Dajani found that the resulting differences in the two approaches for calculating variance estimates for all ATS aggregate totals, all trends of aggregate totals, and approximately 87 percent of the ratios of aggregate totals were not statistically significant. Since the use of replication to calculate the variances of aggregate estimates is easier to program (because covariances do not have to be calculated), Dajani recommended that replication "be used to calculate variances for aggregate estimates in ATS and in other surveys that have a similar sample design."

Though the Census Bureau's economic surveys are primarily single-stage surveys, the Census Bureaus's Survey of Construction (SOC) is a multi-stage, multi-frame survey. The SOC reporting unit is a construction project, not an establishment selected from the Census Bureau's business register. Thompson (1998) and Thompson and Sigman (1998) investigated the use of modified half-sample (MHS) replication to estimate variances for SOC. They recommended that StEPS use VPLX to calculate MHS variance estimates for SOC instead of using legacy code that calculated variance estimates with sampling-theory formulas.

Based on the research studies described above, the Economic Directorate decided in 1997 to calculate sampling variances in StEPS using the following "two methods" approach (Sigman 1997):

! For surveys with single-stage Poisson-sampling designs, use appropriate sampling-theory formulas (see Särndal 1996), and

! For all other survey designs, use one of the replication options of VPLX.

Two recent developments, however, have caused the Economic Directorate to abandon this two-method approach. One of these developments was that following the migration of three surveys into StEPS in 1998 we realized that the implementation effort to use VPLX to calculate random group variances for eleven surveys in 1999 would be very high. The second development was the decision by survey designers of several surveys that in the past had used Poisson sampling to instead use Tillé sampling (Tillé 1996, Slanta 1999). As a result of these two developments, we replaced the two-method approach with the following four-method approach:

! For surveys with single-stage Poisson-sampling designs, use appropriate sampling-theory formulas.;

! For surveys with single-stage Tillé-sampling designs, use sampling-theory formulas described in Tillé (1996) and Slanta (1999);

! For all other survey designs that use the random group method to calculate sampling variances, use SAS macros %rg_var1 and %rg_var2, contained in StEPS (described in section 5.3); and

! For all surveys that do not use the random group method to calculate sampling variances, use a replication option in VPLX.

In section 8 we discuss areas of future research, the findings from which may result in additional changes to the StEPS list of methods for calculating sampling variances.

4. Overview of StEPS

StEPS is a generalized survey processing system that the Economic Directorate has developed to replace 16 legacy systems. In addition to reducing resources needed for system maintenance, one of the StEPS objectives is to shift more processing control to survey analysts. StEPS contains integrated modules for data-collection support (e.g., mail-label printing and questionnaire check-in); editing; data review and correction; imputation; calculation of estimates and variances; and system administration (e.g., parameter specification and the submission and monitoring of batch jobs). Functions not in StEPS include: frame development, sample selection, actual data collection, and dissemination.

StEPS is programmed in SAS, and it stores data and parameters in SAS data sets. The Economic Directorate executes StEPS mainly on Compaq® Alpha® machines using UNIX as the operating system. Most users access StEPS via a graphical (X-Windows) communication package loaded on their desktop microcomputer.

Ahmed and Tasky (1999, 2000) provide additional information StEPS. Tasky et al. (1999) describe the StEPS system design and associated programming strategies. In particular, they state that the developers of StEPS "decided on four major design concepts:

1) "Design a set of standard data structures that remain the same, regardless of the survey and the data.

2) “Use parameters (stored in general data structures) to drive the survey-specific processing requirements.

3) “Generate a ‘fat’ record data set on the fly for certain modules ... .

4) “Standardize field names and possible value for similar concepts.”

The next section describes how these design concepts were implemented in the StEPS Estimates and Variances Module.

5. Components of the StEPS Estimates and Variances Module

In a large survey organization, the specification of survey processing operations requires inputs from multiple specialists. This is especially true when specifying the calculation of estimates and sampling errors. Survey analysts knowWHAT estimates to calculate with WHAT data. The sample designer knows HOW to calculate the estimates and sampling errors. Thus, one of the functional requirements for the StEPS Estimates and Variances Module was that it permit specification of both WHAT and HOW information. A second functional requirement was that it be able to calculate estimates and variances for many different surveys. A third functional requirement was that the Estimates and Variances Module must be integrated with the other StEPS modules–i.e., it should, where possible, use data sets used by other modules; and its interactive screens should have a similar “look and feel” as those for other StEPS modules.

Like other StEPS modules, the following are the major components of the StEPS Estimates and Variances Module:

! Standard data set structures for micro data, macro data, and processing parameters;

! Interactive screens for specifying parameters, submitting batch jobs, and requesting results listings; and

! SAS macros and scripts for batch calculations.

Each of these is discussed below.

5.1. Standard data set structures

StEPS stores micro data in *control files* and *item files*. Micro data includes data associated with questionnaire items; data associated with survey operations such as sample selection, mailing, collection, or check-in; or auxiliary data available from censuses or administrative sources. The item file can contain only numeric micro data, whereas the control file can contain numeric and character data. Another difference between the control file and the item file is that the control file has a “fat” format, whereas the item file has “skinny” format. In the control file (i.e., fat format) there is one record per reporting unit (ID), and the fields within each record correspond to control-file variables. In the item file (i.e., skinny format) there is one record per ID/item combination, and fields within each record correspond to different *data versions* (plus there is a field containing a data flag).

StEPS stores the following data versions in each record of the item file:

r_{ij} = reported data for item i and reporting unit j
e_{ij} = edited data for item it and reporting unit j
a_{ij} = adjusted data for item it and reporting unit j
w_{ij} = weighted-adjusted data for item it and reporting unit j

The default value for edited data is $e_{ij} = r_{ij}$. StEPS users, however, may change edited data by using the Review and Correction Module, or StEPS can change edited data via the Imputation Module.

Some surveys adjust micro data for data collection effects, such as trading day effects in monthly surveys or in annual surveys the effect on reported inventories of ending inventory dates other than December 31. One way that StEPS adjusts micro data is

$$a_{ij} = f(t_i,\boldsymbol{B}_j)\, e_{ij} ,$$

where

t_i = the value for item i of a variable, called *adjustment type* stored on the *item data dictionary* file;
$\boldsymbol{B}_j$ = a vector of *BY variables* --i.e., categorical variables--associated with reporting unit j; and

f( ) = a SAS format that StEPS creates to map the vector $(t_i,\boldsymbol{B}_j)$ into user-provide adjustment factors. Another way StEPS adjusts micro data is to use user-provided SAS code stored in the *adjust/derive definitions file*. Many surveys do not adjust their micro data, however, in which case $a_{ij} = e_{ij}$.

StEPS calculates weighted-adjusted data using the following formula:

$$w_{ij} = ?_j \, g_{n(i)\,,\,j} \, a_{ij} \; .$$

The quantity $?_j$ is the sampling weight for reporting unit j. The control file stores three *g weights*, g_{1i} , g_{2i} , and g_{3i} , for each reporting unit. We had planned to use the g-weights in the manner described in Estavao, et al. (1995), in which if they are chosen appropriately the resulting weighted totals (or weighted means) are generalized regression estimators. To date, we have not used the g-weights for this purpose. One way we have used the g-weights was in our annual retail trade survey, which collects some items for only a subsample of the survey, we let $?_j$ store the first-phase sampling weight and let the g-weight store the second-phase weight. In the future we plan to use the g-weights to store non-response adjustment factors for surveys that use weight adjustment to handle unit nonresponse. The g-weight is equal to 1.0 for unweighted and Horvitz-Thompson estimators. The quantity *n(i)* is the *g-weight number* and indicates which g-weight, g_{1i} , g_{2i} , or g_{3i} , is associated with item i. If *n(i)*=0 then item i has a g-weight of 1.0. The g-weight number, like the adjustment type, t_i, is stored in the item data dictionary, which contains one record for each item-data variable.

The item file's skinny format can be difficult to use for estimation and variance calculations. Consequently, StEPS can create an *estimation fat file*, which has one record per ID, and the fields within each record can be any of the following: control file variable; adjusted or weighted-adjusted version of an item file variable; constant data; or *recode*, which is a variable created at the time of fat-file creation via a user-provided SAS expression involving other fat-file variables. When StEPS creates an estimation fat file, a variable on the control file, called the *weighting switch*, selects for each ID the adjusted or weighted-adjusted version of the item file variables. Certain values of the weighting switch zero out item data in the fat file or delete an entire record from the estimation fat file. By setting the weighting switch to a particular value for each ID, one can control the contents of each estimation-fat-file record, for purposes such as handling deaths by zeroing out or deleting data or handling outliers by deleting or down-weighting to self-representing.

StEPS stores macro data in *estimation results files* (ERFs). One ERF corresponds to one *table*, which is the result of StEPS performing calculations on *analysis variables* for individual values of categorical *BY variables*. The types of results StEPS stores in ERFs include: totals, ratios, trends, other derived estimates (i.e., functions of totals), standard errors, CVs, covariances, t-tests, imputation rates and disclosure-avoidance information. The ERF has a skinny format--each ERF record contains only one calculated result, with other variables in the record identifying the type of result, the name(s) of the analysis variable(s), and the value(s) of any BY variable(s).

Two files store estimation processing information: the *estimation specification file* (ESF) and the *estimation formulas file* (EFF). The ESF stores parameters used by the SAS macros described in section 5.3; the EFF stores SAS expressions and SAS code, also used by the SAS estimation macros. Both the ESF and EFF are populated via interactive screens. Developing a file layout for the ESF was challenging. We rejected a skinny-record format of one record per parameter because of the complexity of file updating from screens displaying multiple parameters. Instead, we decided the ESF would have one record per *specification*, which is a vector of parameters displayed together on the same screen. In the ESF, sets of specifications (i.e., records) associated with the same type of screen and processing action are called *objects*. For example, the "BY object" contains specifications for BY variables, whereas the derived object contains specifications for the calculation of derived estimates.

5.2. Interactive screens

Interactive screen in the Estimates and Variances Module allow StEPS users to do the following:

! Calculate weighted data for all items and IDs in the item file.

! Run Quicktab program, which calculates weighted totals, year-to-year trends, imputation rates, unweighted counts, and disclosure-avoidance information. The Quicktab program requires analysis variables to be item file variables and any BY variables to be control file variables. Quicktab does not calculate standard errors, CVs, or derived estimates. The possible outputs from Quicktab are a SAS data set, an ASCII file (for downloading), printer output, or the SAS Output Window.

! Enter and modify specifications and formulas for use by batch jobs. Specifications and formulas tell StEPS WHAT to estimate with WHAT data. The StEPS user can select analysis and BY variables (from item data, control data, recodes, or constants); specify the method of calculating standard errors (random groups, VPLX replication, or formulas for Poisson or Tillé samplig); enter formulas for derived estimates and the derivatives of non-linear estimators); copy results from one ERF to another ERF; and remove results from an ERF.

! Submit estimation scripts to run in batch. Scripts are described in more detail in section 5.3. A screen displays the available scripts, and the user selects one of the displayed scripts to run immediately or at a scheduled time.

! Review estimation results. A screen displays a list of ERFs, and the user can select an ERF for interactive viewing with SAS/FSVIEW® or for formatting by StEPS into a printed listing.

5.3. SAS macros and scripts

StEPS scripts execute SAS code that is part of StEPS or has been generated by StEPS. For the Estimates and Variances Module, scripts execute SAS code that is part of StEPS. In particular, estimation scripts execute one or more of the following SAS macros:

%extract — Creates estimate fat file.
%totals — Calculates totals and imputation rates.
%derive — Calculates derived estimates.
%erfrmt — Reformats an ERF.
%rtsumvar — Aggregates totals, standard errors, and imputation rates.
%copy1 — Copies results between ERFs.
%remove — Removes results from an ERF.
%round — Rounds totals and standard errors
%vpl2stp — Stores VPLX-calculated estimates and standard errors in an ERF (Dajani 1999a).\
%rgvar_1 — Calculates replicate totals from random group totals.
%rgvar_2 — Calculates replicate-based standard errors from replicate estimates.
%vrncs_p — Calculates standard errors for Poisson-sampling designs.
%vrncs_t — Calculates standard errors for Tillé sampling designs.
%cvrncs_t — Calculates covariances for Tillé sampling designs.
%taylor — Calculates standard errors of non-linar estimates using Taylor approximation.

Many of these macros are individually controlled by parameters analysts have entered into the ESF and EFF. Parameters specify WHAT to estimate and WHAT data to use. The estimation script controls the overall logic of HOW to calculate estimates and variances. Because this depends on the sample design, surveys with different sample designs require different scripts. Also, the sample designer should be involved in developing an estimation script--either as an advisor or as the person who produces the script.

6. Examples of estimation scripts

StEPS has two type of scripts: generic scripts and complete scripts. A *generic script* is a SAS program that executes SAS code contained in StEPS or generated by StEPS. In a generic script, macro variables are used to refer to the survey, statistical period, and other job submission conditions. StEPS users (or StEPS implementation staff) prepare generic scripts using the SAS Editor or other word processing package. A *complete script*, on the other hand, is created by StEPS, and it links to a corresponding generic script (via a %include statement). A complete script defines the macro variables that are used in the corresponding generic script.

Example 1: Create fat file, calculate totals and derived estimates, and calculate CVs using method of random groups.

Generic script:

```
01    *sc_no=F001  sc_desc=Example1;
02     %include ‘/steps/central/autocall.sas’;
03     %setlibs(survey=&survey,statp00=&statp00);
04     %getstime;
05     %let ntbles=1; %let table1=F401;
06     %extract;
07     %totals(rg=y);
08     %rg_var1(intab=F401, outerf=F401);
09     %derive;
10     %rg_var2(inerf=F401,outerf=F401);
11      %applog(module=estimate, submod=&sc_no,
                    starttme=&startme,prgnme=&sc_no,
                    otherinfo=&sc_no);
```

Discussion: Since **line 1** begins with an “*”, it is a SAS comment and is ignored by the batch processing. The information in line 1 is used, however, by the interactive script-submission screen to identify the script number (F001) and the script description (“Example 1"). This information appears on the script-submission screen in the list of scripts available for submission. **Line 2** makes StEPS SAS code available to the batch program. **Line 3** creates all needed SAS LIBNAMEs and UNIX environment variables. **Line 4** puts the starting time into the macro variable &startme. **Lines 5 and 6** create the estimation fat file needed to calculate totals. **Line 7** calculates random group totals and stores the results in ERF F401. **Line 8** converts the random group totals in ERF F401 into replicate totals. **Line 9** updates ERF F401 with derived estimates calculated for each replicate. **Line 10** calculates standard errors from the replicate estimates and stores them in ERF F401 along with the full-sample estimates and the associated CVs. **Line 11** puts information in the production log about the completed batch job.

Example 2 : Calculate totals and imputation rates for XSALES00 and XESALE00 with BY1=state and BY2=NAICS6 (i.e. six digit NAICS code) and store in ERF F301. File ERF F302 contains census totals CSALES with BY1=NAICS2 (i.e., two digit NAICS code). Adjust XESALES00(state,NAICS6) by multiplying by

$$F(NAICS2) = CSALES(NAICS2) / XSALES00(NAICS2) .$$

Generic script:

```
01-04 ... (comment, %setlibs, %getstime)
05     %let ntbles=2;
06     %let table1=F301; %let table2=F302;
07     %extract;
08     %let ntbles=1;
09     %totals(imprate=y);
10     %rtsumvar(intn=F301, outn=F302);
11     %let table1=F302;
12     %derive; ** Calculate F(NAICS2) **;
13     %erfrmt(intn=F302, outn=F301,
                    incndtn=%str(ITEM EQ F));
14     %let table1=F301;
15     %derive; **Calculate adj XESALE00 **;
16     .... (%applog)
```

Discussion: **Lines 1 through 4** are the same as Example 1. **Lines 5. 6 and 7** create an estimation fat file containing all the variables needed to calculate totals (in line 9), aggregate results (in line 10), and reformat an ERF (in line 13). **Lines 8 and 9** calculate totals and imputation rates for ERF F301 with BY1=state and BY2=NAICS6. **Line 10** aggregates results in ERF F301 and puts the aggregated results in ERF F302, with BY1=NAICS2. **Lines 11 and 12** calculate the adjustment factors and stores them in ERF F302. **Line 13** reformats the adjustment factors in ERF F302 into the structure of ERF F301, where the reformatted adjustment factors are stored with BY1=state and BY2=NAICS6. **Lines 14 and 15** calculate adjusted XESALE00 values and updates ERF F301 with these results.

7. Implementation Experiences

In 1998 the Economic Directorate used StEPS for production processing of three annual surveys. The largest of these was the (wholesale) Annual Trade Survey (ATS), with a stratified sample of approximately 7000 reporting units. The other two surveys were small industrial-product surveys; one was a Poisson sample with less that 600 reporting units and the other was a cut-off sample with less that 200 reporting units. In 1998 interactive screens for entering estimation specifications had not yet been developed. Thus, development staff typed estimation parameters into fixed-field ASCII files for the three surveys. This was tedious and error prone, which motivated the development of interactive screens that were used in 1999. In 1998 the StEPS developers created the estimation script files for these three surveys.

In 1998 we used VPLX to calculate variances for ATS using the method of random groups. The length of the VPLX program used to calculate these variances was 513 lines. Since in 1999 there would be an additional ten StEPS surveys using random groups to calculate variances, we concluded that the implementation effort involved in continuing to use VPLX for these surveys was unacceptably high. Hence, we decided to develop the StEPS macros %rg_var1, %rg_var2, %rtsumvar, and %erfrmt for calculating random group variances in 1999.

The production of variances for ATS was improved considerably in 1998 over what was possible from the legacy system. ATS calculates census-adjusted estimates, and the legacy system incorrectly treated the adjustment factors as constants. StEPS, however, was able to correctly calculate the variances of the ATS census-adjusted estimates by treating them as ratio estimates. Another improvement over the legacy system was that the interactive script-submission capability of StEPS permitted survey analysts to obtain estimates and variances upon demand, which was not possible from the legacy system.

In 1999 the Economic Directorate used StEPS for the production processing of fifty surveys. Eleven of these surveys were service-sector surveys that used random groups to calculate variances. These surveys ranged in size from as small as 4,000 reporting units to as large as 27,000 reporting units. Two of the surveys processed by StEPS in 1999 used Poisson sampling; and one survey, the Manufacturing Energy Consumption Survey, used Tillé sampling. The remaining 30+ surveys were industrial-products surveys that used cut-off sampling.

In 1999, survey analysts for the eleven service-sector surveys and for one of the Poisson-sample surveys used interactive screens to enter estimation specifications into StEPS. As in 1998, however, estimation scripts were for the most part created by StEPS development staff. In 1999, we did not use VPLX to calculate random-group variances–instead, we used StEPS macros to calculate random group variances for the eleven service-sector surveys. For the Poisson-sample and Tillé-sample surveys, we will use the StEPS macros %vrncs_p and %vrncs_t, respectively, to calculate variances. For the 30+ surveys that were cut-off samples, analysts used Quicktab to calculate estimates, so it was not necessary for scripts to be written or for estimation specifications to be entered into StEPS.

8. Future Activities

The Economic Directorate plans to migrate additional surveys into StEPS. One of these is the Survey of Construction (SOC), which will use the Estimates and Variances Module of StEPS (and not the other modules in StEPS). Because SOC is a multi-stage survey, we will use VPLX to calculate SOC variances. Another survey migrating into StEPS is the Annual Capital Expenditures Survey (ACES), which has a stratified sample design. We plan to investigate the stratified jackknife for estimating variances for ACES.

The use of standard data structures in StEPS facilitates comparative methodological research. We plan on comparing replication-based variances with those calculated using the sampling-theory formulas for Poisson and Tillé sampling. Another study we plan on conducting will compare random group variances to those from a delete-a-group jackknife.

Finally, we observe that the development of a survey processing system is a journey, not a destination. Lessons learned from today's processing suggest enhancements to the system to perform tomorrow's processing. One possible enhancement to the Estimates and Variances Module is to provide a graphical user interface for developing estimation scripts. Another possible enhancement is to interface estimation results files from StEPS to online analysis tools such as SAS/EIS® and SAS/INSIGHT®.

References

Ahmed, S. and Tasky, D. (1999), "The Standard Economic Processing System: A Generalized Integrated System for Survey Processing, " *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, to appear.

Ahmed, S. and Tasky, D. (2000), "Standardized Economic Processing System," *Proceedings of the International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association, to appear.

Dajani, A. (1999a), *Version 1.0 of %vpl2stp: A SAS macro for Creating SAS Data Sets from VPLX Display-step Output*, Technical Report #ESM-9901, Washington DC: Bureau of the Census.

_______ (1999b), *Comparison of Variance Estimation Methods for Aggregate Estimates*, Technical Report #ESM-9902, Washington, DC: Bureau of the Census.

Estavao, V.; Hidiroglou, M; and Särndal, C. (1995), "Methodological Principles for Generalized Estimation System at Statistics Canada," *Journal of Official Statistics*, **11**, pp. 181-204.

Fay, R.E. (1990), "VPLX: Variance Estimates for Complex Samples," *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, pp. 266-271.

King, C. and Kornbau, M. (1994), *Inventory of Economic Area Statistical Practice, Phase 2: Editing, Imputation, Estimation, and Variance Estimation*, Technical Report #ESMD-9401, Washington DC: Bureau of the Census, March 1994.

Kott, P. (1999), "Some Problems and Solutions with a Delete-A-Group Jackknife", *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, Tuesday B Sessions, Washington, DC: Council of Professional Associations on Federal Statistics, pp. 129-135.

Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics*, **1**, pp. 381-397.

Särndal, C. (1996), "Efficient Estimators with Simple Variance in Unequal Probability Sampling," *Journal of the American Statistical Association*, **91**, pp. 1289-1300.

Sigman, R. (1997), "How Should We Proceed to Develop Generalized Software for Survey Processing Operations Such as Editing, Imputation, etc?", unpublished paper presented at the Meeting of Census Advisory Committee of Professional Associations, Washington DC: U.S. Bureau of the Census, May 1-2, 1997.

Slanta, J. (1999), "Implementation of Modified Tillé Sampling Procedure in the MECS and R&D Surveys,"*Proceedings of the Survey Research Section*, Alexandria, VA: American Statistical Association, to appear.

Tasky, D.; Linonis, A.; Ankers, S; Hallam, D., Altmayer, L.; and Chew, D. (1999), "Get in Step with StEPS: Standard Economic Processing System," *Proceedings of the North East SAS Users Group*, pp. 167-178.

Thompson, K. (1998), *Evaluation of Modified Half-Sample Replication for Estimating Variances for the Survey of Construction (SOC)*, Technical Report ESM-9801, Washington DC: Bureau of the Census.

__________ and Sigman, R. (1998), "Modified Half Sample Variance Estimation for Median Sales Prices of Sold Houses: Effects of Data Grouping Methods," *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, pp. 698-703.

Tillé, Y (1996), "An Elimination Procedure for Unequal Probability Sampling Without Replacement," *Biometika*, **83**, pp. 238-241.

Town, G. (1997), *The Use of VPLX to Calculate Variances for Economic Surveys Using the Method of Random Groups: A Tale of Two Surveys*, Technical Report #ESM-9701, Washington DC: Bureau of the Census.

Tremblay, T. (1996), "Estimates from VPLX Random Groups Option," unpublished memorandum, Washington DC: Bureau of the Census, September 26, 1996.

__________ and Sigman, R. (1996), "Comparison of Two Variance-Estimation Methods for a Standardized Economic Processing System,"*Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, pp. 204-208.

Wolter, K. (1985), *Introduction to Variance Estimation*, New-York: Springer-Verlag.